ESD-TR-67-380

MTR-413

# ALTERNATIVES IN EVALUATION OF COMPUTER SYSTEMS

Dennis W. Fife

DECEMBER 1968

Prepared for

DIRECTORATE OF PLANNING AND TECHNOLOGY

ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts

Project 512B
Prepared by
THE MITRE CORPORATION
Bedford, Massachusetts
Contract AF19(628)-5165

ESD-TR-67-380 MTR-413

# ALTERNATIVES IN EVALUATION OF COMPUTER SYSTEMS

Dennis W. Fife

DECEMBER 1968

Prepared for

DIRECTORATE OF PLANNING AND TECHNOLOGY

ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts

Project 512B
Prepared by
THE MITRE CORPORATION
Bedford, Massachusetts
Contract AF19(628)-5165

## FOREWORD

Contractor: The MITRE Corporation
Bedford, Massachusetts 01730

Contract Number: AF19(628)-5165

Air Force Contract Monitor: Charles L. Bruce, ESLFA

This report covers research aimed at recognizing, collecting, and disseminating significant data for evaluation purposes. The effort was performed between 23 February 1967 and 1 April 1967.

## REVIEW AND APPROVAL

This technical report has been reviewed and is approved.

WILLIAM F. HEISLER, COL, USAF
Chief, Command Systems Division
Directorate of Planning & Technology

# ABSTRACT

This essay is a commentary on the formal conduct of computer evaluation studies. Two contrasting viewpoints are discussed — that of a computer specialist and that of a non-specialist concerned with system acquisition. The opinion expressed is that any evaluation task normally has unique aspects which demand the attention and creative efforts of a computer specialist. For the specialist, a prescribed evaluation form and an associated scoring procedure, as frequently used in acquisition activities, do not constitute a satisfactory methodology for a technical evaluation. They are useful for planning, or as a means of documentation if suitably modified and extended during the evaluation effort.

## TABLE OF CONTENTS

# SECTION I

# INTRODUCTION

Most people in the computer field do some amount of evaluation work in an effort to keep up with the current trends, research studies, and commercial products in the field. The results are informal opinions, largely based on experience and casual reading, and usually rendered in bull sessions or at conferences. Once in a while, however, a more formal evaluation task comes along, expressed in terms such as "evaluate system X for data management and information retrieval." The investigation and conclusions in this case are to be rendered in a formal document. Formal evaluation activities, certainly a concern in the past, are bound to arise more often as the spectrum of commercially-available hardware and software broadens. So it seems worthwhile to seek a common viewpoint on the nature of these projects, the means of accomplishing them, and the current needs in evaluation studies.

This paper presents a brief perspective on these matters, focusing on the question of how mechanical and routine one can make an evaluation activity. There are advocates of a universal and routine method for evaluation of computer systems, especially among those faced with equipment purchasing decisions. In such a method, the investigation is conducted and documented by filling out a prescribed chart, matrix, or tabular form containing entries which are deemed appropriate and sufficient for a wide spectrum of evaluation tasks. However, the constraint to follow such a form and fill in every line may be more burdensome and diverting than helpful. Instead, the practical and unique circumstances of any evaluation demand a flexible and evolutionary evaluation technique. The following discussion hopefully will evoke some agreement with the notion that research on prescribed, general purpose evaluation forms is less desirable than research aimed at recognizing, collecting, and disseminating significant data for evaluation purposes.

# SECTION II

## THE GENERAL APPROACH TO EVALUATION

It is fair to say that a formal evaluation task, whether dealing with computers or not, may have one of two purposes: to provide the technical basis for an impending decision, or to provide defensible justification for a decision that has already been made. The latter implies a prior bias and will not be considered further, but it often occurs. There are three steps then to the accomplishment of an evaluation study. Paraphrasing Markel(1), one first determines a set of questions or subject areas to be addressed by the evaluation. The basis for doing this is the decision which motivates the evaluation. Second, one collects data appropriate to each question or subject. Third, one forms a judgement based on the evaluation data collected.

The process of formulating the evaluation questions is crucial. It establishes, first of all, an organization and study discipline for the evaluation project. It involves decisions on what is pertinent to the impending decision that has motivated the evaluation. It includes some judgement of the relative importance of features and distinctions between systems, reflected at least in the amount and direction of the initial evaluation effort.

The particular circumstances of an evaluation have a substantial bearing upon the proper evaluation questions. Part of the circumstances are historical and deal with past evaluation studies, established applications requirements, and current experience with the systems being studied. Even more important are the intended duration and resources available for the evaluation task. The latter may establish by default how comprehensive, discriminating, and technically substantial an evaluation can be.

In any case, an evaluation should always produce evidence of sound technical consideration, which leads to the matter of collecting evaluation data. There are three sources of data for evaluation purposes: experiments and usage experience; simulation, analysis, and thought experiments; published data and solicited observations. Experiments and usage experience involve direct contact with the physical subject of evaluation. This may be as simple as a programming problem used to get a feel for a computer language. Or, it may involve a complex scenario with people, computers, sensors, precise schedules of events, and elaborate measurements. Jacobs(2) describes some SAGE experiments of this nature. The advantage of experiments, when honestly attempted, is that the greatest realism short of prolonged operational experience is attainable. The disadvantages are, of course, cost and the difficulty of experimental control.

Simulation, analysis, and thought experiments do not, in my view, deal with the physical system directly but rather with a model or abstraction of the system. The model may be described in a computer program, as in the usual Monte Carlo processes, or in terms of equations, when the analysis is mathematical. One advantage of the use of models is simplification, with the possibility of saving evaluation costs and achieving clear insight into system behavior. In many cases, models allow the investigation of behavior that is practically impossible to treat experimentally, thus implying flexibility as another advantage. The disadvantages of the modeling approach are the problems of validating the underlying assumptions and the limited applicability of existing models.

Published data and solicited observations arise, of course, from reference manuals, papers, interviews, and questionnaires. The advantage of these sources is that the data is available and relatively easy to obtain. The disadvantages are that the data may be poorly organized, insufficient, inaccurate, or irrelevant for the evaluation task.

Finally it is worthwhile to comment on the process of forming judgements about a system. Jacobs (2) has given a useful categorization of the alternative viewpoints which evaluators may take in considering the value or merits of a system. Jacobs distinguishes four attitudes, termed respectively the excellence, utility, desirability, and formality orientations toward a system's value. The excellence approach is typified by a techniques researcher, e.g., a specialist in sorting algorithms, who judges a system by a performance measure such as average sort time per item. The excellence-oriented evaluator wants to achieve the most desirable performance. In the case of the utility viewpoint, typically taken by a system engineer, the evaluator is less concerned with optimum performance in one technical area than with meeting stated requirements of an application which involves many technical areas. The desirability notion is assumed by managers and military commanders, who must consider cost and with limited resources choose systems to fulfill a number of jobs. The formality approach applies to acquisition managers and purchasing agents who are often principally concerned that a system pass prescribed regulations such as a standards specification.

This categorization indicates that a technically stronger but less formalized and routine evaluation would be expected from those of the excellence and utility viewpoints. It also implies that when an opinion is offered, the justifications are based on comparison.

The comparison may be with the state of the art, the capabilities of competing systems, the requirements of an application, the needs established with past usage experience, or the opinions of authorities.

Formality-oriented evaluators, however, have developed additional techniques of judgement, involving numerical scores or figures-of-merit, that are presumably to provide more objectivity or to avoid any initial bias toward one of the competing alternatives. Such techniques will be examined shortly. At this point in the history of machine computing there is little objective data on application requirements and system trade-offs, so it hardly seems possible to be objective in judgement. The subjective evaluations given by informed and experienced professionals are a necessary and unavoidable aspect of present-day evaluation projects.

# SECTION III

## THE TECHNICIAN'S METHODOLOGY

As suggested by Jacobs, evaluators with primarily technical interests are chiefly concerned with performance measures for a system. The key ingredients in establishing and predicting performance are empirical data and analysis. Empirical data, obtained through benchmark problems or operational experience, may be used to justify performance criteria, to verify conclusions reached by analysis, and to establish approximations and parameter values used in system models. Analysis of models may be used to extrapolate from simple empirical observations and thereby estimate performance in situations which cannot economically be tested physically.

It is surprising, in view of the scientific foundations of the computer field, that so little effort is made to apply mathematical analysis or to undertake empirical observations which lead to clear understanding of the performance consequences of design decisions. However, there is no reward in minimizing the reasons this situation exists. Mathematical analysis is not especially relevant in some aspects of design, such as judging the ease of using system capabilities. In areas where it is clearly useful, such as production performance in terms of throughput and response time, not enough background yet exists to make its application routine. Moreover, not enough readily-available empirical data exists, nor do manufacturers provide hardware and software which facilitate empirical measurements of an application environment. The digital clocks provided on computers in the past have been unstable or have had insufficient resolution, and it is expensive to fit software into manufacturer-provided operating systems and translators in order to realize a tracing or measurement function.

Calingaert(3) provides a useful survey of the current situation in performance evaluation. Some examples of mathematical analysis are Scherr(4), Smith(5), Fife(6), and Coffman(7). McIsaac(8), and Belady(9) treat simulation models. Examples of the environmental statistics necessary are given by Scherr, Rosin(10), and Irani, et al(11).

## SECTION IV

## THE FORMALIST'S METHODOLOGY

By far the largest portion of published material explicitly dealing with evaluation as a formal activity treats procedures used in acquisition or purchasing, especially of hardware. (See reference (12) for example). The basic technique is to compute a single numerical score or figure-of-merit for each competing alternative in a decision, and choose the alternative having the highest score. The score is developed from a list of system attributes appropriate to the application. This list is to be established before knowing the competing alternatives in the impending decision. A weight is assigned for each attribute according to its "importance", again prior to any knowledge of the available alternative systems. For each alternative system, measurements or observations are made of its attributes and a score is assigned which depends upon the observed or measured value. Thus, one uses an equation:

$$\text{System score} = \sum_{i=1}^{N} \alpha_i f_i(x_i) \qquad (1)$$

where

$N$ = number of applicable attributes,

$\alpha_i$ = weight for i<u>th</u> attribute,

$x_i$ = observed or measured system value for i<u>th</u> attribute,

$f_i$ = scoring function for i<u>th</u> attribute, having observed value as its argument.

Miller(13) has considered this method of decision-making in some depth, and has carried out an experiment to assess its merit. The approach rests, of course, on the postulate that a preference ranking of alternative systems is realizable via the numerical score computed for each one. Moreover, for expediency in computation, it is assumed that the score may be obtained by adding independent contributions due to individual system attributes. Complicated interrelationships among the attributes are therefore neglected in deducing preference. Regarding equation (1) then, the proposed attributes must be such that the weights, $\alpha_i$, can be assigned without regard to observed or measured values of any of the attributes. Unfortunately, rather little guidance is provided toward accomplishing this or recognizing when it has been achieved. Apparently it is a very intuitive and subjective process and, as Miller finds, rather difficult for experimental subjects to handle properly.

It should be emphasized that Miller does not view this technique as producing increased objectivity in a decision. Instead his results indicate that it induces greater personal confidence in a subjective judgement. Miller found, for example, that the experimental subjects would suggest modifications to the admissible attributes, weights, or scores whenever the computed preference ordering did not conform to their subjective judgement of what it should be. The computation was thus used to substantiate or clarify the basis for a prior subjective evaluation.

An extensive attribute list developed for use in such a method is given in a report of Informatics, Incorporated(14). The list is pertinent to general purpose data management systems. The system parameters are organized into functional areas, such as "data definition", "file generation", and "retrieval". One area, termed "environment", encompasses computer hardware considerations, installation management costs and resources, and other factors which are not easily associated with any one functional area. Under each category, the parameters are organized into subcategories such as "file definition", "file security", and "editing". Typical single parameters are "files identifiable by name", "protection of file against accidental update", and "suppression of leading zeros on output". Altogether there are in the neighborhood of 500 individual capabilities and parameters listed.

Even so, an evaluator will need to examine the list carefully, discarding some attributes, adding other attributes, or expanding upon the description given. For example, capabilities involved in graphical input-output and time-shared operation are listed, and these may not be relevant in a particular evaluation effort. The collection of attributes must also be studied with a view toward achieving the independence of criteria required by the scoring technique. Thus substantial intuitive and subjective work is needed to achieve a suitable parameter list, even with such extensive raw material.

But remember that in the formalists' approach this effort is supposedly carried out for an evaluation task before any knowledge of the specific alternatives to be evaluated has been obtained. The evaluator, however, is bound to make intuitive assumptions about how the attributes will be satisfied,which may not hold true when the actual alternatives are presented. This creates the necessity for adding and changing attributes and weights after seeing what capabilities are available. For example, one might require simply that a programming language should allow complex variables, and assume in so stating that it will of course allow one to iterate on a complex variable. The language PL/I, however, allows complex variables but not in an iteration statement(15). Thus the evaluator's assumptions may not be valid, and this casts serious doubt on whether an evaluation form can be properly completed without studying the available alternatives.

A prescribed, extensive evaluation form may become largely a distraction and a burden in an evaluation task. It is quite unlikely that it will be precisely suited to a particular task. The process of manipulating an extensive list forces the evaluator to devote time to marginally significant criteria, thus giving less time to pursue the important criteria in depth. The formalists' requirement that the form is to be complete and weights irrevocably assigned before studying the available alternatives does not recognize the fact that intuitive assumptions of how capabilities are supposed to be satisfied may not be valid. Finally, the numerical manipulations of scoring and weighting create an additional burden whose contribution is very questionable in view of the highly subjective effort which precedes it.

Lists of parameters or criteria, such as contained in references (14), (16), (17), and (18), can nonethless be very useful as initial guides in formulating evaluation questions and organizing a study. By giving increased confidence that no important area is neglected entirely, they can contribute to a sense of objectivity and reliability in the conclusions. Questionnaires moreover are a valuable device for collecting and documenting evaluation data. However, the evaluator should be free to direct the evaluation effort according to what his experience and growing knowledge of the competing systems indicates are the crucial factors. He should not be constrained to follow a general purpose approach which does not account for the unique and unforeseen circumstances of a particular task.

## SECTION V

## CONCLUSION

Two things are clear about evaluation tasks. An evaluator must be informed and experienced, and must attempt to understand and determine the requirements of the application, or context, for the evaluation. Further, the approach for the evaluation must be suited to the constraints which apply to the task. These constraints will limit the extent of the evaluation effort, the feasible amount of data collection, the scope of experiments, etc. Thus it seems self-defeating at the present time to espouse a single routine methodology for computer system evaluation.

The need for benchmark application problems seems common to all evaluation philosophies, including that of the formalists. From a practical standpoint, even a specialist is likely to be unfamiliar with the details of a particular system to be evaluated. Benchmark problems are thus a means of learning the system, and a focus for the necessary familiarization effort. Suitable benchmark problems should therefore be developed for any application as a means of testing proposed system capabilities. They can also, as simple experiments, provide rudimentary performance data. The latter can be extrapolated to explore performance limits by means of simulation or mathematical models.

The most prominent need in evaluation work is empirical data, on application environments and on actual performance of computer hardware and software. Collection and dissemination of data will provide a reasonable basis for establishing requirements and factors which are truly significant to an evaluation. Empirical data is also a needed input in formulating assumptions and simplifications for the creation of system models. Research along these lines should be supported because formal understanding is the only eventual avenue to a reliable general purpose evaluation approach.

## REFERENCES

1. Markel, G.A., Toward a General Methodology for Systems Evaluation, HRB Singer Company, Report 352-R-13, (DDC AD619 373), Contract Nonr 3818(00), State College, Pennsylvania, July 1965.

2. Jacobs, J.F., Practical Evaluation of Command and Control Systems, MTP-7, MITRE Corporation, Bedford, Massachusetts, November 1965.

3. Calingaert, P., "System Performance Evaluation: Survey and Appraisal," COMM. ACM, 10, January 1967, 12-18.

4. Scherr, A., An Analysis of Time-Shared Computer Systems, Massachusetts Institute of Technology, TR-18, Project MAC, Cambridge, Massachusetts, June 1965.

5. Smith, J.L., "An Analysis of Time-Sharing Computer Systems Using Markov Models," in PROC. SJCC, Spartan Books, Incorporated, Washington, D.C., 1966, 87-96.

6. Fife, D.W., "An Optimization Model for Time-Sharing," in PROC. SJCC, Spartan Books, Incorporated, Washington, D.C., 1966, 97-104.

7. Coffman, E.G., Stochastic Models of Multiple and Time-Shared Computer Operations, Department of Engineering, UCLA, Report No. 66-38, Los Angeles, California, June 1966.

8. McIsaac, P.V., Job Descriptions and Scheduling in the SDC Q-32 Time-Sharing System, System Development Corporation, TM2996, (AD 636 839), Santa Monica, California, 10 June 1966.

9. Belady, L.A., "A Study of Replacement Algorithms for a Virtual Storage Computer," IBM Systems Journal, 5, 1966, 78-101.

10. Rosin, R., "Determining a Computer Center Environment," COMM. ACM, 8, July 1965, 463-468.

11. Irani, K., et al., A Study of Information Flow in Multiple Computer and Multiple Console Data Processing Systems, Rome Air Development Center, RADC-TR-65-532, Rome, New York, New York, February 1966.

12. Rosenthal, S., "Analytical Technique for Automatic Data Processing Equipment Acquisition," in PROC. SJCC, Spartan Books, Incorporated, Washington, D.C., 1964, 359-381.

13. Miller, J.R., The Assessment of Worth: A Systematic Procedure and Its Experimental Validation, Massachusetts Institute of Technology, Ph.D. dissertation, Cambridge, Massachusetts, 1966.

14. Buettell, T., et al., A Methodology for Comparison of Generalized Data Management Systems: PEGS (Parametric Evaluation of Generalized Systems), Informatics, Incorporated, TR-66-655-8, Sherman Oaks.

15. IBM Corporation, PL/I Language Specifications, IBM System Reference Library, Form C28-6571-4, New York, December 1966.

16. Davis, R., "Programming Language Processors," Advances in Computers, Vol. 7, Academic Press, New York, New York, 117-180, 1966.

17. Shaw, C.J., An Outline/Questionnaire for Describing and Evaluating Procedure Oriented Programming Languages and Their Compilers, System Development Corporation, FN-6821/000/00, Santa Monica, California, August 1962.

18. Budd, A., A Method for the Evaluation of Software, ESD-TR-66-113 Vols. 1, 2, and 3, The MITRE Corporation, Bedford, Mass., July 1967.

## DOCUMENT CONTROL DATA - R & D

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

| | |
|---|---|
| 1. ORIGINATING ACTIVITY *(Corporate author)*<br>The MITRE Corporation<br>Bedford, Mass. | 2a. REPORT SECURITY CLASSIFICATION<br>UNCLASSIFIED<br>2b. GROUP |

3. REPORT TITLE

ALTERNATIVES IN EVALUATION OF COMPUTER SYSTEMS

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

N/A

5. AUTHOR(S) *(First name, middle initial, last name)*

Dennis W. Fife

| | | |
|---|---|---|
| 6. REPORT DATE<br>December 1968 | 7a. TOTAL NO. OF PAGES<br>16 | 7b. NO. OF REFS<br>18 |

| | |
|---|---|
| 8a. CONTRACT OR GRANT NO.<br>AF 19 (628) - 5165 | 9a. ORIGINATOR'S REPORT NUMBER(S)<br>ESD-TR-67-380 |
| b. PROJECT NO.<br>512B | |
| c. | 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*<br>MTR-413 |
| d. | |

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

| | |
|---|---|
| 11. SUPPLEMENTARY NOTES<br>N/A | 12. SPONSORING MILITARY ACTIVITY Directorate of Planning and Technology, Electronic Systems Division, Air Force Systems Command, USAF, L. G. Hanscom Field, Bedford, Massachusetts |

13. ABSTRACT

This essay is a commentary on the formal conduct of computer evaluation studies. Two contrasting viewpoints are discussed - that of a computer specialist and that of a non-specialist concerned with system acquisition. The opinion expressed is that any evaluation task normally has unique aspects which demand the attention and creative efforts of a computer specialist. For the specialist, a prescribed evaluation form and an associated scoring procedure, as frequently used in acquisition activities, do not constitute a satisfactory methodology for a technical evaluation. They are useful for planning, or as a means of documentation if suitably modified and extended during the evaluation effort.

| 14. KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| EVALUATION TECHNIQUES | | | | | | |
| EQUIPMENT SELECTION | | | | | | |